УДК 681.3

О.О. Марченко
Київський національний університет імені Тараса Шевченка
факультет кібернетики
Україна, 03680, м. Київ, просп. Глушкова, 4-д

СИСТЕМА АНАЛІЗУ КОРЕФЕРЕНТНИХ ЗВ'ЯЗКІВ У ТЕКСТАХ

O.O. Marchenko
Taras Shevchenko National University of Kyiv
Faculty of Cybernetics
Ukraine, 03680, Kyiv, Glushkova Ave., 4-d

SYSTEM FOR COREFERENCE ANALYSIS OF NATURAL LANGUAGE TEXTS

А.А. Марченко
Киевский национальный университет имени Тараса Шевченко
факультет кибернетики
Украина, 03680, г. Киев, просп. Глушкова 4-д

СИСТЕМА АНАЛИЗА КОРЕФЕРЕНТНЫХ СВЯЗЕЙ В ТЕКСТАХ

У статті описана побудована система аналізу кореферентних зв'язків у текстах природною мовою. У роботі були реалізовані два алгоритмічні підходи для моделювання кореферентних відношень тексту та машинного навчання системи визначення та аналізу зв'язків – із застосуванням методу максимальної ентропії та з використанням методу опорних векторів.
Ключові слова: обробка текстів природною мовою, кореферентний аналіз, семантичний аналіз

The article describes a developed system for co-reference analysis of natural language texts. Two algorithmic approaches for modeling coreferential relations in natural language texts and for training the system for determination and analysis of relationships – Maximum Entropy and Support Vector Machine were used.
Keywords: natural language text processing, co-reference analysis, semantic analysis

В статье описана разработанная система анализа кореферентных связей в текстах на естественном языке. В работе были реализованы два алгоритмических подхода для моделирования кореферентных отношений в тексте и машинного обучения системы определения и анализа связей – с применением метода максимальной энтропии и с использованием метода опорных векторов.
Ключевые слова: обработка текстов на естественном языке, кореферентный анализ, семантический анализ

Вступ

Система аналізу кореферентних зв'язків у текстах була розроблена на кафедрі математичної інформатики факультету кібернетики Київського національного університету імені Тараса Шевченка. Вона призначена для аналізу англомовних текстів. Її основною задачею є визначення всіх сутностей тексту та встановлення для кожної групи іменника тексту на яку саме сутність даний іменник посилається. Зазначимо, що дана задача повністю розв'язується лише на рівні семантичного аналізу.

Основна складність кореферентного аналізу випливає із фундаментальної проблеми мовної полісемії. З одного боку, одну і ту саму сутність у тексті можна виразити багатьма різними способами, а, з іншого боку, завдяки омонімії одне й те саме слово у різних місцях тексту може посилатися на різні сутності. Тому встановлення однозначних зв'язків між іменниками та сутностями тексту є складною проблемою, яка на сьогоднішній день є відкритою, і розв'язання якої вимагає значних зусиль по створенню потужних систем семантичного аналізу природної мови на основі великих онтологічних баз знань. Кореферентний аналіз включає у себе також таку складну та фундаментальну для комп'ютерної лінгвістики задачу, як розв'язання займенникової анафори. Як і для будь-якого іншого іменника, для займенників у тексті також треба знайти їх антецеденти (іменники, на які вони посилаються) і вказати для них відповідні їм сутності тексту.

Побудована система кореферентного аналізу реалізована із застосуванням таких потужних підходів машинного навчання як метод максимальної ентропії та метод опорних векторів. Дані методи добре зарекомендували себе, зокрема, у комп'ютерній лінгвістиці, як точні методи обчислення розв'язку для різноманітних задач класифікації. Розроблені моделі кореферентних зв'язків у текстах природною мовою дозволили системі отримати показники точності аналізу, що на стандартних тестових корпусах переважають відомі найкращі світові аналоги.

Архітектура системи

Структурно система представляє собою послідовність блоків аналізу текстів, кожен з яких послідовно здійснює аналіз тексту. Речення тексту обробляються послідовно. На першому етапі роботи система за допомогою блоку морфологічного аналізу виконує лексико-морфологічний аналіз речення та визначає для кожного слова його нормальну форму та морфологічні характеристики (частина мови, рід, число, відмінок, час і т.д.). Наступним етапом аналізу є синтаксичний аналіз, який виконується блоком синтаксичного аналізу, що за даними попереднього етапу обробки вибудовує дерево підпорядкування (dependency tree) речення. Далі дані передаються на блок кореферентного аналізу, що розв'язує наступну задачу:

Дано: текст англійською мовою, що пройшов етапи лексико-морфологічного та синтаксичного аналізу.

Знайти: обчислити список сутностей E_1, E_2, E_3,…, E_n, що згадуються у тексті. Для кожної групи іменника NP_i вказати таку сутність E_k, на яку NP_i посилається.

Задача блоку кореферентного аналізу полягає у побудові розбиття – треба розбити множину NP на класи E. Елементи множини NP мають набір властивостей – морфологічних, синтаксичних та семантичних. Система має виконати кластеризацію множини іменників у тексті, застосовуючи принципи відповідності семантичного, синтаксичного та морфологічного рівня. У процесі роботи блоку кореферентного аналізу поточний NP_i має бути зарахований до одного з відомих класів сутностей тексту E_1, E_2, E_3,…, E_{k-1}, якщо буде мати місце відповідність з елементами відповідного класу, або для поточного NP_i буде заведено новий клас сутності тексту E_k у тому випадку, якщо відповідності немає – тобто сутність нова і вище по тексту не згадувалася. Задача визначення класу E_k для NP_i розв'язується наступним чином.

NP_i розглядається як анафора, для якої потрібно знайти антецедент у попередніх реченнях тексту. Якщо дана задача буде виконана, то можна зарахувати NP_i до класу E_k, до якого належить відповідний NP_i антецедент.

Для розв'язання даної задачі використовується ряд фільтрів для значного зменшення кількості класів-кандидатів E. Фільтри першого рівня – морфологічні. Вони передбачають відповідність анафори та антецеденту за такими характеристиками як рід, число і т.д. Якщо у деяких кандидатів у антецеденти має місце невідповідність, то їх можна відразу відкинути разом з їх класами сутностей E. Фільтри другого рівня_– синтаксичні. Вони містять набір правил синтаксичної відповідності анафори та антецеденту, такі як синтаксичний паралелізм, несумісність анафори та антецеденту як аргументів одного дієслова і т.д. Використовуючи синтаксичні фільтри, також можливо відчутно скоротити кількість кандидатів у антецеденти та відповідні класи-кандидати E. Фільтри третього рівня – семантичні. Вони представляють собою процедури перевірки на семантичну відповідність іменника NP_i та іменників, що належать класам сутностей E. Якщо вони належать до різних семантичних класів, то відповідні класи-кандидати E відкидаються. Процедури перевірки побудовані на основі лексико-семантичної бази WordNet [1] із застосуванням алгоритмів обчислення міри семантичної близькості між словами [2].

Після застосування різнорівневих фільтрів залишається незначна кількість кандидатів у антецеденти і, відповідно, незначна кількість класів-кандидатів E. Далі блок кореферентного аналізу виконує задачу класифікації NP_i серед класів сутностей E, що не були відкинуті в процесі фільтрації, із застосуванням моделі максимальної ентропії [3] та методу опорних векторів [4]. Для цих методів окремо були сформовані вектори ознак для NP_i та кандидатів-антецедентів. Ці вектори ознак містили семантичну, синтаксичну та морфологічну складові. Окремо для кожного методу розв'язувалася оптимізаційна задача підбору найкращого набору ознак для векторів, який би відповідав максимальним оцінкам точності роботи блоку кореферентного аналізу на тестових корпусах.

Модель максимальної ентропії

Основа ідея, закладена в метод Максимальної Ентропії, полягає у тому, що використовуються лише наявні дані і не створюється жодних припущень щодо розподілення ймовірностей над даними, які не є присутніми в системі. Дана модель відноситься до класу *умовних* або дискримінантних імовірнісних моделей, що є найбільш широко застосовуваним у вирішенні задач з комп'ютерної лінгвістики, розпізнання мовлення та у машинному навчанні взагалі.

Основними перевагами моделі є:

- висока точність;
- дозволяє легко працювати з лінгвістично важливими ознаками (властивостями);
- дозволяє будувати мовнонезалежні моделі для вирішення різних задач комп'ютерної лінгвістики.

Дискримінативні моделі вираховують імовірності $P(c|d)$ прихованих структур, спираючись на вхідні навчальні дані без передобробки, тобто моделюється лише умовна ймовірність класів.

Ознаки. Ознака f – елементарна частина певної ознаки, що пов'язує дані d, які ми розглядаємо, з категорією C, яку ми передбачаємо для цих даних. Модель визначає вагу кожної ознаки наступним чином:

- позитивну вагу, якщо ознака ймовірно вірна;
- негативну вагу, якщо ознака ймовірно невірна.

У проекті використовувалися булеві значення ознак (yes/no), як це прийнято в комп'ютерній лінгвістиці. Кожна ознака «обирає» підмножину даних і пропонує для неї мітку.

На етапі класифікації система виконує наступні дії:

1. Обирається лінійна функція з набору ознак $\{f_i\}$ в класи $\{C\}$.
2. Встановлюються ваги λ_i для кожної ознаки f_i.
3. Кожні дані d, що представляють собою певні NP, перевіряються на належність до кожного з класів C.
4. Для кожної пари (c,d) ознаки голосують з урахуванням своїх ваг: vote (c) = $\Sigma\lambda_i f_i(c,d)$.
5. Обирається той клас, що максимізує $\Sigma\ \lambda_i f_i(c,d)$.

Сама модель Максимальної Ентропії (MaxEnt) не є новою, проте застосування її до задачі вирішення кореференцій є досить нестандартним і потребувало розробки нових ідей. Основна проблема лежить у тому, що MaxEnt створювався і зараз використовується як класифікатор, тобто він здатен розділити на класи подану на вхід множину слів або документів. На перший погляд здається, що цього достатньо: необхідно прийняти як класи сутності E, а як елементи – групи іменника NP. Проте, це рішення вдало підходить для даної задачі лише на перший погляд: потенціальна кількість таких класів у нас є нескінченою, і, навіть, якщо ми зможемо якось обійти дане обмеження і створити набори ознак для кожного з класів, залишиться проблема навчання. Система буде здатна розпізнати і зібрати як класи лише ті сутності E, що були присутні в навчальному корпусі, і пропустить всі, що є новими для неї. Природно, що така ситуація не є прийнятною і потребує нових підходів для вирішення.

Основною ідеєю, застосованою в рамках реалізації адаптованої до задачі моделі максимальної ентропії, є пропозиція як класи розглядати лише два класи, один умовно можна назвати «Кореферентні», інший − «Некореферентні». Як елементи треба використовувати не самі NP, а їх пари. Використовуючи дану ідею, класифікатор ділить усі вхідні дані на два класи (що дає суттєвий виграш у швидкості, порівняно з моделлю мультикласифікації) і, що набагато важливіше, достатньо побудувати досить обмежений набір ознак, необхідних для запуску класифікатора. Фактично, для початку роботи класифікатора достатньо мати одну булеву ознаку, позитивне значення якої трактується як належність до класу «Кореферентні», а негативне – до класу «Некореферентні».

Навчання та тестування проводилося на основі корпусів розмічених текстів Ontonotes [5]. У текстах корпусу вручну проставлені мітки кореферентних зв'язків між NP. На основі текстів корпусу була отримана навчальна вибірка потрібного об'єму та вмісту. У результаті експериментів було підібрано оптимальний набір

ознак для векторів NP, що відповідає найкращим показникам точності роботи блоку визначення кореферентних зв'язків між NP.

Метод опорних векторів

Основна ідея методу опорних векторів (SVM) – перетворення вхідних векторів у простір більш високої розмірності, де є висока ймовірність, що дані будуть лінійно-роздільними, та пошук роздільної гіперплощини з максимальним зазором у цьому просторі. У випадку задачі знаходження кореферентних відношень вхідні вектори формуються у вигляді прикладів відповідно до обраної моделі представлення (пари NP, кластер-NP, ранжування NP чи кластерне ранжування) та складаються з виділених ознак. Дві паралельні гіперплощини будуються по обидві сторони гіперплощини, яка розділяє дані класи. *Роздільною гіперплощиною* буде гіперплощина, яка максимізує відстань до двох паралельних гіперплощин. При цьому таку гіперплощину називають *оптимальною гіперплощиною*, а точки даних, які лежать ближче всього до цієї гіперплощини, називаються *опорними векторами*. Алгоритм працює таким чином, що чим більша відстань буде між паралельними гіперплощинами, тим менша середня помилка класифікатора.

Класифікація на основі SVM полягає в наступному:

1) Вхідні вектори подаються на вхід SVM у вигляді прикладів відповідно до обраної моделі представлення та складаються з виділених ознак.
2) Бажані значення d_i (вчитель) – це значення, що характеризують кореферентність: 0 або 1 (або значення рангу, якщо використовується у моделях ранжування NP та кластерного ранжування).
3) Навчання SVM базується на розв'язанні задачі квадратичного програмування з використанням методу множників Лагранжа.
4) Для випадку лінійної нероздільності у цільову функцію замість скалярних добутків вводиться нелінійна функція ядра.

У результаті навчання SVM отримуємо оптимальний вектор вагових коефіцієнтів, що визначає перпендикуляр до роздільної гіперплощини та оптимальне значення порогу. Знайдені параметри підставляються у рівняння гіперплощини для нових точок, таким чином здійснюється класифікація. Як програмна реалізація методу опорних векторів був використаний програмний пакет SVM-Light.

Експерименти

Для навчання та тестування були використані корпуси розмічених текстів Ontonotes. А саме, Wall Street Journal Corp., Newswire, Broadcast News, Web text. Навчання та подальше тестування проводилося по методу Cross Validation, згідно з якими весь корпус ділиться на N частин, система навчається на N−1 частинах корпусу, а тестування проводиться на одній частині, на якій не проходило навчання. Потім обирається інша наступна одна тестова частина корпусу із зсувом на одну позицію вліво, навчання системи відбувається на решті корпусу, а тестування на новообраній частині. Так циклічно відбувається N сесій навчання-тестування, під час якого тестова частина проходить через весь корпус і кожна з N частин даних використовується для тестування. У результаті отримаємо оцінку ефективності моделі з найбільш рівномірним використанням наявних даних. Із отриманих N оцінок точності можна фіксувати мінімальні значення, та розглядати

їх як гарантовані оцінки точності роботи системи. Під час тестування обчислювалися точність роботи системи кореферентного аналізу (P, precision), повнота (R, recall) та інтегрована оцінка F:

$$P = \frac{кількість_правильно_визначених_зв'язків}{число_усіх_знайдених_кореферентних_зв'язків};$$

$$R = \frac{кількість_правильно_визначених_зв'язків}{число_усіх_кореферентних_зв'язків_тексту};$$

$$F = \frac{2PR}{P+R}.$$

Отримані в результаті оцінки можна побачити у таблиці 1. Для порівняння у таблиці 1 в третьому стовпчику надано оцінки роботи системи кореферентного аналізу, розробленої у Стенфордському університеті (Stanford Deterministic Coreference Resolution System) [6]. Дана програма, на сьогоднішній день, є одним з найкращих світових аналогів побудованої системи кореферентного аналізу текстів.

Таблиця 1. Оцінки роботи системи кореферентного аналізу текстів

	Метод максимальної ентропії	Метод опорних векторів	Stanford Deterministic Coreference Resolution System
P (точність)	79.64	85.00	62.4
R (повнота)	84.34	86.00	59.3
F	81.89	85.49	60.8

Як можна побачити, оцінки системи кореферентного аналізу, отримані під час експериментів переважають показники Стенфордської системи. Особливо високими виявилися оцінки у блока, реалізованого на основі моделі опорних векторів (SVM). Проте, слід відмітити, що навчання даного блоку вимагає набагато більше часу порівняно з моделлю максимальної ентропії.

Висновки

У роботі представлено опис нової розробленої системи аналізу кореферентних зв'язків у текстах природною мовою. Була запропонована оригінальна архітектура системи, яка суміщає різні рівні лінгвістичного аналізу тексту, послідовну семантичну, синтаксичну, морфологічну фільтрацію, що значно скорочує кількість варіантів при визначенні зв'язків типу *група іменника-сутність* та блоки ідентифікації та аналізу кореферентних зв'язків, реалізовані із застосуванням двох основних підходів до машинного навчання – моделі максимальної ентропії та методу опорних векторів. Експерименти з розміченими текстовими корпусами показали високу точність роботи системи кореферентного аналізу на рівні кращих світових аналогів.

Література

1. Miller G. A., Beckwith R., Fellbaum C. D., Gross D., Miller K. WordNet: An online lexical database // International Journal of Lexicography – 1990. – 3, 4. – pp. 235-244.

2. Marchenko O. O. Methods for Estimations of Semantic Closeness-Relatedness of Natural Language Words // Artificial Intelligence. – 2012. – 4. – pp. 213-219.
3. Berger A.L., Pietra V.J.D., Pietra S.A.D. A maximum entropy approach to natural language processing // Computational Linguistics (MIT Press). – 1996. – 22 (1). – pp. 39-71.
4. William H., Teukolsky S. A., Vetterling W. T., Flannery B. P. "Section 16.5. Support Vector Machines". Numerical Recipes: The Art of Scientific Computing (3rd ed.). New York: Cambridge University Press. – 2007. – ISBN 978-0-521-88068-8.
5. URL – https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf
6. Lee H., Chang A., Peirsman Y., Chambers N., Surdeanu M., Jurafsky D. Deterministic coreference resolution based on entitycentric, precision-ranked rules // Computational Linguistics. – 2013. – 39(4). – pp. 885-916.

RESUME

O.O. Marchenko
System for Coreference Analysis of Natural Language Texts

The paper presents the description of a new developed system for coreference analysis of natural language texts. The problem of coreference analysis of natural language texts is one of the classical and fundamental tasks of computational linguistics. The coreference analysis of natural language texts consists in definition of all entities for some input natural language text and solving the problem of finding correct corresponding entity for each noun phrase of the text. It should be noted that the problem of coreference analysis can be completely solved only on the semantic level of natural language text structure with applying special semantic analysis algorithms on the base of ontological knowledge bases.

The paper describes an original system architecture which contains multilevel linguistic analysis of text, successive semantic, syntactic, morphological filtering, that greatly reduces the number of options in determining relationships such as "Noun - Entity" and subsystems for identification and analysis of coreference relations that were implemented by using two major approaches to machine learning – maximum entropy and support vector machines.

Experiments with text corpora showed high accuracy of coreference analysis. Support vector machine demonstrates higher estimates of precision, recall and accordingly F-measure than the maximum entropy method values. However, it should be noted that the training of the support vector machine method requires a much longer time than the maximum entropy model.

The system estimates overcome the best world analogies.

О.О. Марченко
Система аналізу кореферентних зв'язків у текстах

У роботі представлено опис нової розробленої системи аналізу кореферентних зв’язків у текстах природною мовою. Аналіз кореферентних зв’язків у текстах природною мовою є одною з класичних та фундаментальних задач комп’ютерної лінгвістики. Кореферентний аналіз природномовного тексту полягає у визначенні всіх сутностей вхідного тексту та у розв’язанні задачі

знайдення коректної відповідної сутності для кожної групи іменника у тексті. Слід відзначити що задача кореферентного аналізу може бути повністю розв'язана лише на семантичному рівні структури тексту із застосуванням спеціальних алгоритмів семантичного аналізу на основі онтологічних баз знань.

Стаття описує оригінальну архітектуру системи, яка суміщає різні рівні лінгвістичного аналізу тексту, послідовну семантичну, синтаксичну, морфологічну фільтрацію, що значно скорочує кількість варіантів при визначенні зв'язків типу *група іменника-сутність* та блоки ідентифікації та аналізу кореферентних зв'язків, реалізовані із застосуванням двох основних підходів до машинного навчання – моделі максимальної ентропії та методу опорних векторів.

Експерименти з розміченими текстовими корпусами показали високу точність роботи системи кореферентного аналізу. Метод опорних векторів демонструє вищі оцінки точності, повноти та, відповідно, F-міри, ніж метод максимальної ентропії. Але слід відзначити, що навчання методу опорних векторів вимагає набагато більше часу, ніж у моделі максимальної ентропії.

Показники роботи системи на рівні кращих світових аналогів.

Надійшла до редакції 28.08.2015.